Computational Analysis of Merchant Marine GPS Data*

CASOS Technical Report

George B. Davis, Kathleen M. Carley

November, 2006
CMU-ISRI-07-109

Carnegie Mellon University
School of Computer Science
ISRI - Institute for Software Research International
CASOS - Center for Computational Analysis of Social and Organizational Systems

Abstract

A series of quantitative and structural analyses are applied to geospatial data regarding the movement of Merchant Marine vessels in the English Channel.

Keywords: Geospatial analysis, network analysis, clustering

* This work was supported in part by the National Science Foundation under the IGERT program, 9972762, for training and research in CASOS and the Office of Naval Research N00014-02-1-0973. Additional support was provided by CASOS - the center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied of the National Science Foundation, the Office of Naval Research, or the U.S. government.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Computational Analysis of Merchant Marine GPS Data	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **School of Computer Science, Carnegie Mellon University,Pittsburgh,PA,15213-3891**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
A series of quantitative and structural analyses are applied to geospatial data regarding the movement of Merchant Marine vessels in the English Channel.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**22**	

Computational Analysis of Merchant Marine GPS Data
EXECUTIVE SUMMARY

George B. Davis and Kathleen M. Carley
Computational Analysis of Social and Organizational Systems (CASOS) Laboratory
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh PA 15213
{gbd, carley}@cs.cmu.edu

CASOS has been tasked with developing new computational techniques for analyzing data about Merchant Marine behavior under a Social Network Analysis framework. In this paper we describe an experiment doing so for geospatial data from AIS transponders in 1700+ Merchant Marine vessels during a 5 day exercise in the English Channel. Our analysis has three phases:

1. Spatial clustering algorithms are used to detect places of interest and relationships between entities in the data.
2. Extracted relational information is analyzed in network form. A suite of network analytic measures are applied to find patterns on the network and individual node levels.
3. We apply an intervention analysis which models an intervention (surveying ships at ports) and suggests a strategy for allocating surveillance.

The analysis framework is unusual in taking a relational perspective to spatial data, and novel in its principled treatment of the relationship between spatial, two-mode, and one-mode network representations of data, and in its approach to proposing intervention strategy.

KEY RESULTS

- Our clustering approach finds compelling locations of interest, including some not explained by map data. 60% of predictions were within 5km of visible port infrastructure, 21% were waiting patterns outside of busy ports, 19% were new unknown locations with compelling support in data.

- The Place → Place network is more densely connected, yet the Ship → Ship network is more efficient, supporting shorter paths for the exchange of information or resources.

- We are able to use node-level network measures to identify ships and places with several types of significance: some are busiest, but others are more crucial to the connectivity and efficiency of the network.

- We show how network analytic approaches to selecting ports for increased surveillance can balance the tradeoff between cost / inconvenience of surveillance and informational benefit.

- Several new software tools were developed to facilitate this analysis; they are overviewed in an appendix.

I. Table of contents

II. Index of Tables

II. Index of Figures

1. Introduction

From the 25th to 30th of June 2005, a sensor network queried Automated Identification System (AIS) transponders on merchant marine vessels conducting exercises in the English Channel, recording navigational details such as current latitude and longitude, heading, speed, reported destination, and several forms of identifying information. In total, movements of over 1700 vessels were recorded, with activities ranging from simple shipping lane traversals to apparently complex itineraries with stops at multiple ports of call.

The reasons for the collection of the data are primarily security related. The global shipping system plays a prominent role in a variety of terrorist attack scenarios, both in the United States and abroad: in any country, the ports are both the most likely means of entry for bombs and other weapons, and themselves a prime economic and symbolic target. In addition to being an attractive target, ports are currently considered unsecure – for example, it has been suggested that only 3% of shipping containers entering the United States are directly inspected by customs officials. The sheer volume of commerce conducted via international shipping makes naïve attempts at greater security infeasible, as neither the direct costs associated with detailed surveillance nor the indirect costs incurred by reducing industry efficiency are easily absorbed. If automated techniques such as those designed above can give insight into the behavioral patterns and structural features of the merchant marine population, then limited budgets for surveillance and interdictions can be more precisely targeted to have the greatest impact on overall security. The data under analysis here is especially promising as it represents the result of a relative inexpensive, passive, and consensual surveillance effort.

The data accumulated presents a variety of analytical opportunities and challenges. As a complex and varied set of geospatial paths (as well as other dynamically changing variables), the data beg summary via the application of data mining and knowledge representation techniques. As behavioral data, we can consider patterns in ship movements to be the results of decisions made by professional commanders highly constrained by the high cost of maneuvering and maintaining these huge ships. Finally, the data encodes networks of relationships – such as those between ships, from ships to ports, and from ships to countries, as well as the traces of many other unobserved factors. These networks have their own structural properties which can be probed for a greater understanding of the dynamics of the system.

This paper has two primary goals. The first is a rendering of as much information as possible regarding merchant marine networks and behavioral patterns on the basis of the data given. The patterns detected should inform future research efforts to better understand the community. The second purpose is the assessment of the tools and techniques applied as potential parts of an analysis regime which should be repeated on data gathered in the future.

The paper is organized as follows. In section 2, we provide a brief background on the merchant marine community and on the technical details of our data and the way in which it was accumulated. In section 3, we conduct an analysis of the geospatial aspect of the data, first qualitatively and then by using spatial data mining techniques to infer "points-of-interest" around which various merchant behaviors cluster. In section 4, we extract relational networks from the data and analyze their structural properties using network analysis techniques. The key goals in that section are to identify ships and ports which hold important positions in the relational

network and to analyze topological features of the network overall. Then, in section 5 we conduct an intervention analysis in which we compare two policies for efficient surveillance of ships. We conclude in section 6 by summarizing the patterns we have detected in the data and advantages and disadvantages of the techniques applied, as well as outlining specific future work in progress. Section 6 provides an appendix describing the tools developed during this study.

2. Background

Ships exceeding a certain size or carrying certain cargo types are required in US Coastal waters and many international ports to operate a piece of equipment known as an Automated Identification System (AIS). The AIS is a transponder which implements a communication protocol whereby authorities on land and other ships can query local ships for identification and navigation information. In general, the AIS is directly connected to a Global Positioning System (GPS) and other ship navigational computers, allowing it to automatically generate an accurate report of the vessel's current condition. Table 1 lists fields that were included in the reports analyzed in this study. Note that AIS is a general purpose ocean traffic monitoring protocol, and includes many capabilities not discussed in this paper.

Field	Notes
Tracking Number	Unique identifier assigned by querier to ship
Time	
Time Zone	
Latitude	
Longitude	Measurements accurate to 1'
Sensor	Always 'AIS' (could potentially encode other sources)
Course	Directional heading
Speed	
Nav	Navigational status string (e.g. "UNDERWAY", "MOORED"), apparently (due to typos and nonstandard messages) user-inputted
Destination	Apparently nonstandard field – often blank, sometimes names of cities, specific docks, or other information
Name	Ship's name, user inputted (Many ships apparently change reported name; some use captain's or owner's name instead)
Category	In our case always 'MER' for merchant vessels
Force Class	In our case always '18' for merchant vessels
Flag	Country code for ship's nation of origin
Callsign	Radio identifier for ship
IMO	Unique Int'l Maratime Organization identifier
MMSI	Unique "Mobile Maratime Service Identity", used for automatic parsing of radio messages

Table 1: Fields in Merchant Marine AIS Response

Requests for reports can be targeted to individual ships, or broadcast as a request for all local ships to report navigational information. Coast guards and port operators use regular polling of location information to maintain real-time maps of local traffic. The dataset we analyzed includes 42869 AIS reports from approximately 1729 distinct vessels, over a large geographic range that suggests multiple polling stations. Figure 1 shows the locations of all AIS reports in

their geographic context[2]. The precise borders of the data distribution suggest that it is a selected subset from a larger surveillance database. Large gaps without observations suggest that either certain areas are not traveled, that sensors were not placed in those areas, or that they were omitted from the dataset.

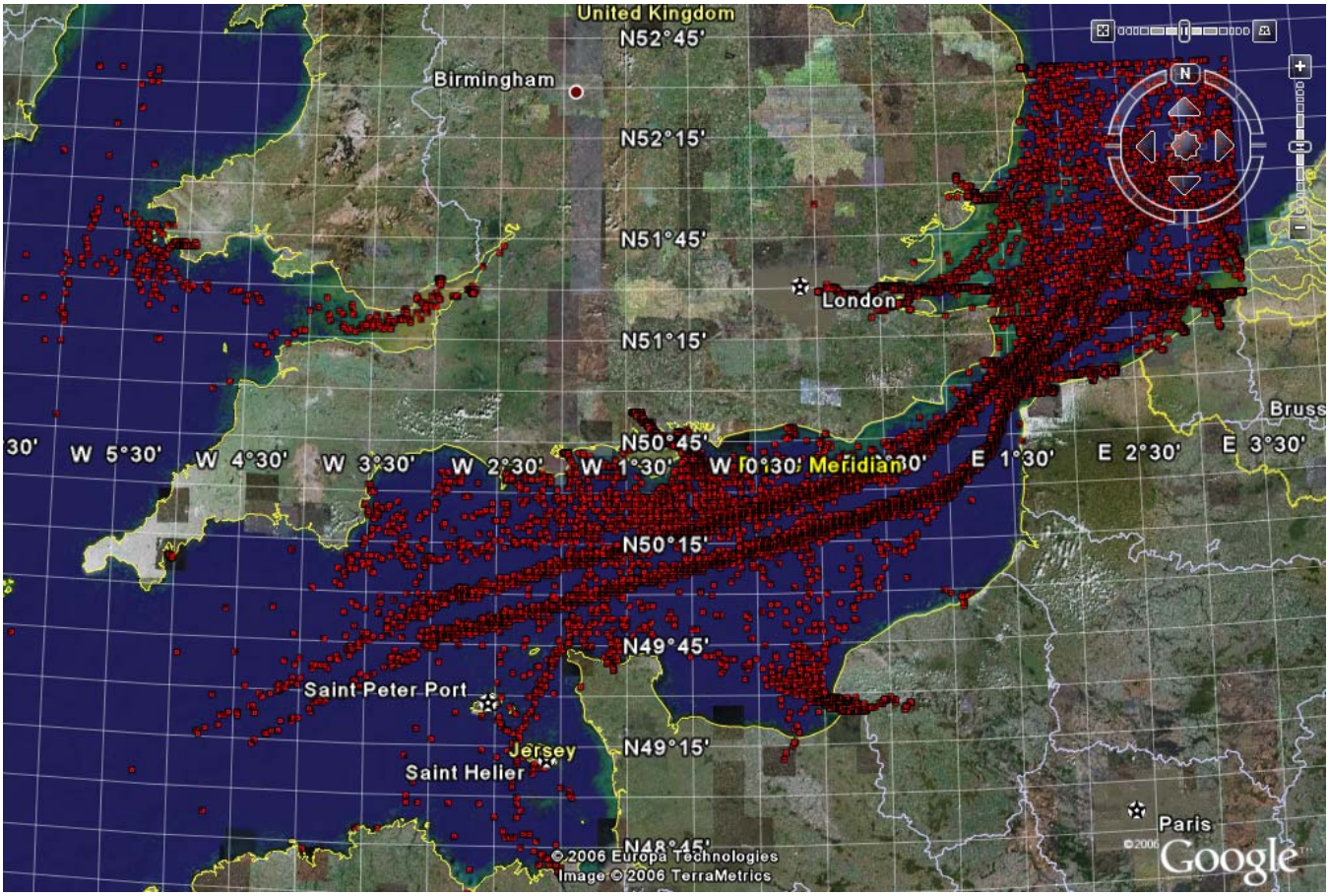


Figure 1. Geospatial distribution of collected data

Although the message format is standardized, several factors prevent consistent and precise interpretation of AIS reports. The precision of the positional fields is fixed but the spacing of the latitude / longitude grid varies around the globe, resulting in position readings that are more precise in some places than others. In the English Channel area, the effective sensor resolution was approximately 1100 meters, or .6 nautical miles, meaning that smaller differences in location could not be accurately distinguished. This means that the data contains no information about behaviors evidenced by more precise movement patterns, such (potentially) usage of different cranes at the same dock. Another form of sensor resolution is the polling frequency and duration. At any given point in the sampled space, queries appeared to be conducted on approximately 40 minute intervals, meaning that activities on a similar timescale might be unrecorded or almost impossible to identify. Additionally, the data we recorded took place over

only 5 days (June 25 – 30, 2005), meaning that it was confined to a specific seasonal context and does not demonstrate much about long-term patterns of behavior.

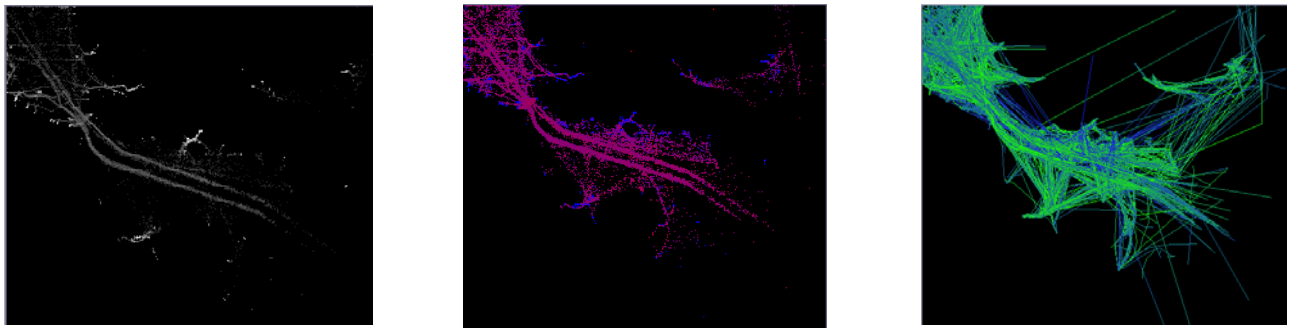
Another source of error in the data is varying standards regarding the installation of equipment and usage of user-specified information fields. For example, we saw many cases in which ships reported 0 velocity even while moving significantly between queries, and others which reported movement while remaining stationary (perhaps they were measuring effort against a local current). Ships were relatively consistent in their use (or lack of use) of unique identifiers such as MMSI and IMO codes, but sometimes would report varying ship names between locations, suggesting that there may be cases of intentional aliasing within the community. Some of the potentially most interesting fields, such as "Destination", were used in many varying ways, making data difficult to interpret.

In this study we focused only on data explicitly recorded from AIS queries. However, opportunities exist to augment this data with other publicly available sources. In particular, the unique MMSI and IMO identifiers, as well as ship's name and nation of origin, offer the potential to match with industry-specific databases regarding ownership and usage of vessels. In addition, it might be possible to infer attributes such as ship size and load from the acceleration characteristics and navigational range of the vessels. This is a promising area for future investigation.

3. Spatial Analysis

3.1 Qualitative Trends

A cursory examination of Figure 1 suggests that the data points are distributed non-randomly, but that visualization does not facilitate any real understanding of the distribution. We designed several other visualizations which provided more traction for the human eye to pick out patterns. Figure 2A is an observation density map, where a pixel's brightness corresponds to the number of observations recorded in the vicinity near the point corresponded to by the pixel center. Figure 2B is a velocity map, where locations with no observations appear black and color scales from blue to red corresponding to low and high average travel velocity nearby all other pixels. Note that in figures 2A and 2B intensity is log-scaled to better express apparently exponential distributions of observation density and average speed. Figure 2C shows trails where points corresponding to the same ship from consecutive time periods are connected to reveal travel paths. Observations near the beginning of the sample period are marked blue, fading to green over time to show direction of travel.



A. Observation Density **B.** Average Speed **C.** Ship Paths

Figure 2. Exploratory visualizations of AIS Geospatial Data

From the density map (2A) one can make out the coastal outline and inland waterways, with bright (high traffic) spots which are presumably ports. Ships at sea seem to be mostly constrained in their trajectories to two primary shipping lanes, with a network of less traveled but nonetheless well defined paths connecting ports to shipping lanes, often along relatively straight paths (i.e. a consistent heading). The average speed map (2B) permits clear distinction between ports of call where ships actually stop and high traffic travel routes. Closer inspection reveals that some of the non-major shipping lanes report much higher than average velocities, suggesting either that faster ships travel these routes or that high traffic is a constraining factor in the main lanes. Finally, by examining the ship paths over time (2C) we can see that many ships traveled all the way across the sample area during the sample period, and that travel occurred in both directions along most routes. Interestingly, some port areas, such as the southernmost and southwestern regions, seem to show a significant net inflow or outflow of ships in the allotted time period, suggesting that we might see cyclical patterns with a longer observation time. Discontinuities in paths, sometimes appearing as straight line jumps over landmass, illustrate that some ships entered and left surveillance. This suggests that either that their paths deviated significantly from those we see (such as taking a northern route around Great Britain) or that their AIS transponder was not active at all times.

3.2 Data Mining

Based on the exploratory observations in 3.1, we framed a data mining goal as follows. Can we extract a set of locations which are "points of interest" around which navigation decisions are made? Note that for any published class of locations – for example, commercial docks or refueling stations – we could accomplish the same task by matching observations against the very databases used by merchant vessel crews to plan their routes. However, inferring these locations directly from data allows us to develop a behavioral model for merchant marines without making assumptions about the types of sites they are likely to visit. One advantage of this is that it prepares us to potentially transfer our analytic techniques to domains where foci of behavior are not well known. Another is that it may better prepare us to decode deviant behaviors that don't fit into our expected profile.

We treated this as a clustering problem, applying the widely used **k-Means** algorithm. K-Means is a *supervised* clustering technique, meaning that machine clustering is preceded by a human analyst selecting a number of clusters and "priming" by specifying initial cluster centers for the algorithm to refine. There exist methods for automating both of these human inputs, but they are beyond the scope of this paper. CASOS is currently working on adapting a more robust and fully automated clustering algorithm, the Conditional Random Field model of Liao *et al.* (2005). We compared the results of k-Means to an "expert" dataset consisting of known ports and refueling stations. We were interested both in the ability of the algorithms to reproduce the known locations of interest and in their identification of previously unknown points, so we examined in detail each point which did not have a match in the database.

k-Means derives its name from the fact that it models the data as coming from *k* collections, each normally distributed around some mean point. One advantage of the algorithm is that various distance metrics can be supplied to find different types of clusters. In our study we used simple geographic distance from some central location, but a more complex model might consider other

ship attributes such as country of origin when selecting clusters. The algorithm proceeds by first choosing some initial means, and then iteratively alternating between assigning observations to the closest mean and improving the location of each mean to best fit the observations assigned to it. Since it is a local search algorithm, it is susceptible to local maxima which could prevent discovery of the true best-fit clusters. For this reason, we had a human analyst manually pick starting cluster locations based on observation density, so that the algorithm's contribution was a refinement of the points he identified.

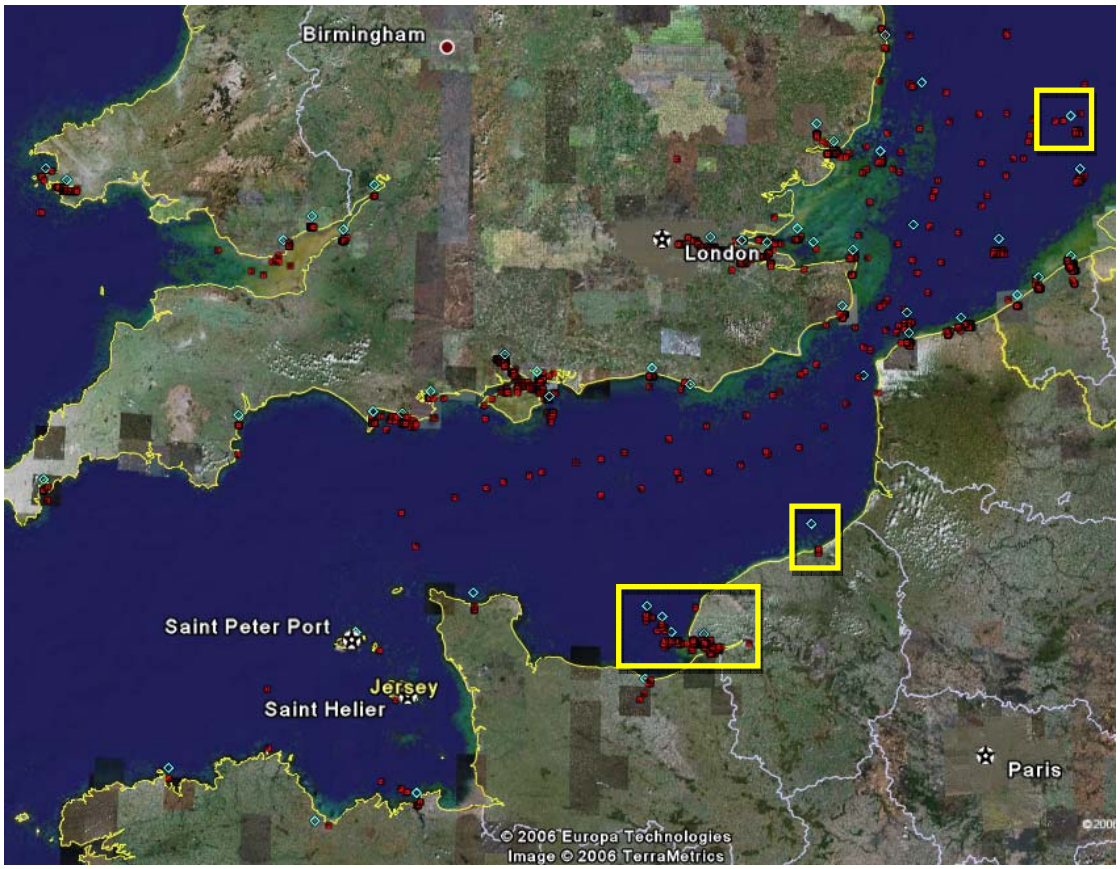


Figure 3. Ships reporting speed < .5 knots (red) and inferred locations of interest (blue).

Figure 3 shows the outcome of the K-Means clustering technique. We compared each reported location of interest to a set of available map data including port coordinates and satellite imagery, and divided the ports into 3 categories.

Direct hits were inferred locations that were within 5 km of clearly visible shipping infrastructure or port coordinates. The 5km cutoff was chosen based on the low resolution of the sensors and the large size of some dock infrastructure. Over half (58.8%) of the locations fell into this category.

Vicinity points were those that were clearly associated with a significant port, but fell outside of the 5 km radius. These comprised 21.5% of our predictions, and could be further classified into two interesting subcategories. At some busy ports, especially those with obstructed entries,

significant clusters of ships could be seen in what might be a waiting pattern outside of the port. The busiest port in the dataset was Le Havre, highlighted in the lowest and leftmost of the yellow boxes in figure 3. Here, several distinct clusters formed a queue leading out of the port and into the channel. 63% of the vicinity predictions fell into this category. The second group were clusters that obviously missed their mark by a significant margin. An example is shown by the second (middle) yellow box: although ships are clearly observed at port, the inferred location is pulled out to sea because it is considered the best explanation for the points erroneously reported as stopped in the shipping lane. We expect cleaner data or an improved clustering algorithm can nearly eliminate this type of error.

Unexplained points were those for which no explanation could be found in our data. An example is given in the top-right yellow box in Figure 3, where a number of ships can be seen clustered at an otherwise non-extraordinary point deep at sea. 19.6% of our predictions fell into this category. We gave these positions special scrutiny, and were interested to note that none of the 10 locations were supported by fewer than 30 observations, with a minimum of 8 distinct ships involved, raising their credibility as genuine foci of behavior.

In final analysis, we considered only 4 of our predictions (7.8%) to be misleading as identifying foci of behavior, and even these clearly corresponded to clusters of activity but were simply rendered inaccurate by noisy data. With an improved clustering algorithm and cleaner data, we feel that even very large spatial datasets could be accurately and automatically annotated with foci of interest, on which further relational analyses can be performed.

4. Structural Analysis

4.1 From Spatial to Structural Data

In relation to the rest of our study, it is useful to view the data mining process in section 3 as a type of noise reduction. Of the many locations we observed for each ship, only a few were selected as intentional destinations, while the rest were dictated by chance elements and the need to abide by conventions and constraints of ocean travel. In our data mining, we leveraged two assumptions -- that points of interest would attract many different ships, and that ships would slow near their intended destinations – to pick out the few locations that ships actively intended to visit.

Collapsing the spatial data into a few decision points in this manner allows us to now consider agent behavior from a relational perspective, asking questions such as “which ships visited which places?”, “which ports were visited by the same set of ships?”, and so on into more complex queries. Since our “ship visiting place” relation is recorded over time in the data, it would be possible to examine the relations dynamically. However, in this paper we will focus on the static set of all relations observed in the time frame, which is captured by the two mode matrix shown in figure 4. In the rest of this paper, we refer to this as the Ship → Place or “Stopped At” network.

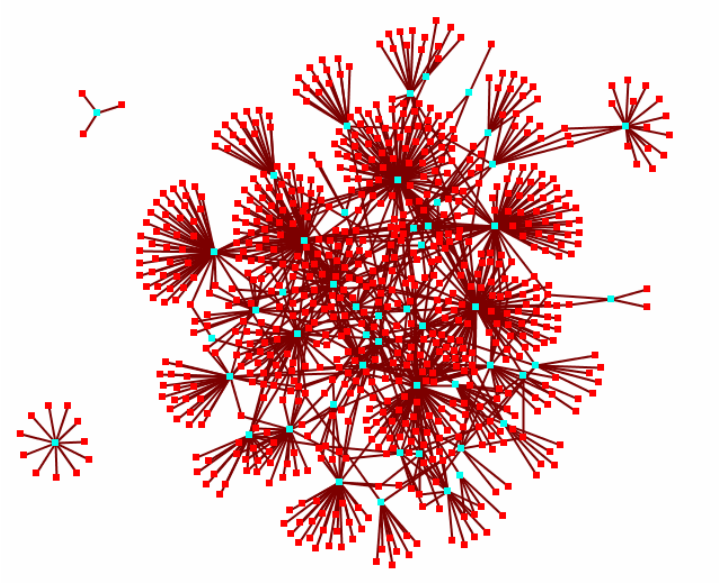
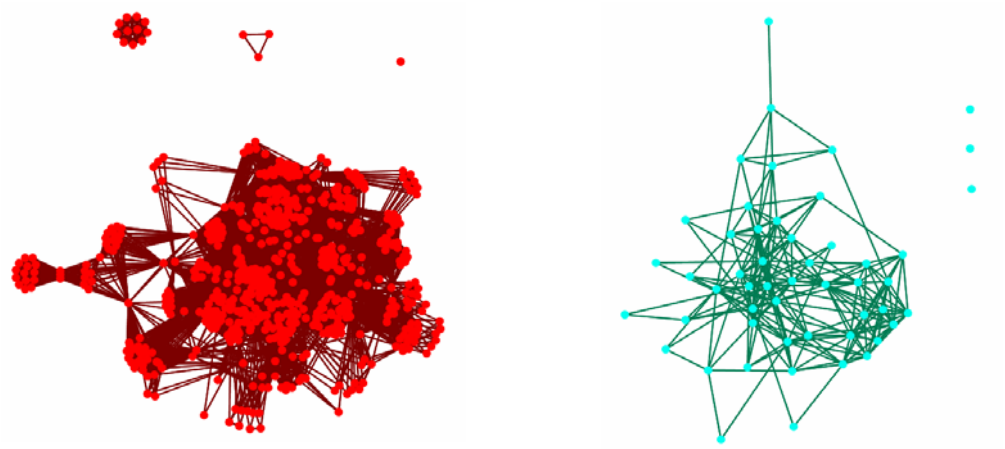


Figure 4. Agent (red) x Port (blue) Network of "Stopped At" Relation

Another interesting class of networks captures compound relationships derived from the primitive relationships shown above. Figure 5 shows two single mode networks derived from the StoppedAt network, where an edge indicates that the two share a neighbor in the bipartite network in figure 4. One way to think of this relation is that neighboring ports can be reached by a single ship route, and that neighboring ships could trade cargo by dropping it at a single port. This type of relationship is sometimes referred to as an algebraic relationship or "word" because it can be calculated by multiplying matrices representing other relationships. For the rest of the paper, we refer to the ship graph as the Shared Port network and the port graph as the Shared Ship network.



A. Shared Port (Ship x Ship) **B.** Shared Ship (Port x Port)

Figure 5. Derived Networks

We can add much of the remaining AIS data to this graph in the form of additional node types and relations. A multi-mode network, also referred to as a meta-matrix, would include nodes representing the various aliases reported by each ship and nodes representing countries with links to ships indicating origination and links to ports representing national territory. Linking to external data sources, such as records of ship ownership, could provide additional entity types with which to enrich this structural model. Multimode networks are as robust as relational databases in their ability to represent data, but it is organized to aide investigation of structural features involving multiple relationships, whereas commercial relational databases are designed to more conveniently investigate distributions of entity attributes under simple relational constraints.

In this paper, we examine the Stopped At, Shared Port, and Shared Ship networks primarily because it is simpler to analyze one and two mode graphs. However, study of more complex networks such as the metamatrix in figure 5 is an important and growing branch of network analysis, and an important area for future investigation of this data.

4.2 Graph Level Properties

Identifying global properties of a network is an important first step in network analysis, since patterns identified here can influence the interpretation of grouping and node-level measures. For example, the presence of several high centrality individuals is unexpected and potentially unstable in a hierarchical network, where one would expect clearly tiered leadership. The same result would be completely typical in a cellular network, where each cell and cell boundary holds influential individuals.

Measure	Shared Ship (Location x Location)	Shared Port (Ship x Ship)	Ship x Port
Nodes	51	749	800
Edges	454	46726	1060
Density	0.178	0.085	0.028*
Clustering Coefficient	0.619	0.891	0.956
Connectedness	0.885	0.961	0.991
Efficiency	0.834	0.916	0.999

Table 2. Standard Unimode Graph-Level Measures.

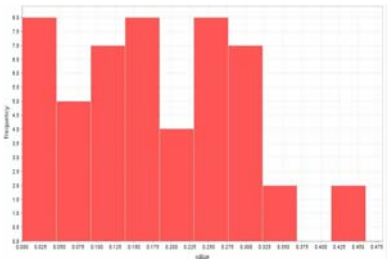
(* adjusted to reflect maximum density of bipartite graph)

Table2 records a series of standard graph-level measures, calculated through ORA's Social Network Analysis report, for three single-mode matrices: the two derived matrices "Shared Ship" and "Shared Port" described in the previous section, and a "unimoded" version of the Stopped At network where ships and ports are interpreted as the same entities. The last of these is included mostly for illustrative purposes, as we will discuss in this section the complications of using unimode network measurements on networks described from two-mode relationships.

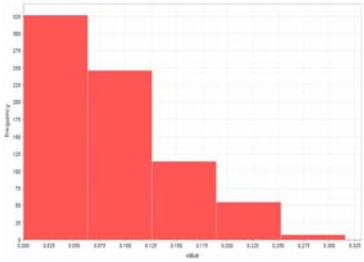
Each measurement is normalized against the maximum possible measurement for a network of the same size. Although these measures are widely applied to any unimode matrix, it is important to remember that both of our matrices were derived from a single two-mode relation. This permits some opportunities for comparison between the graphs. For example, the higher density (fraction of possible edges which exist) on the port network indicates the pattern you might imagine: each ship services a small part of the network within this timeframe, but the aggregate effect of the merchant fleet is that the port network is highly connected. Note that the original bipartite graph has significantly lower density and higher efficiency than either of the derived graphs. This demonstrates leverage, in that the network of relationships actually managed by human decisions (the bipartite network) generates a much richer network of capabilities (the derived networks). The high clustering coefficient in the ship graph suggests that a ship is much more likely to find other ships with similar behavior patterns than a port is to find other ports visited by a similar array of ships.

The high connectedness in both graphs is unsurprising because each was constructed out of a series of cliques – for example, in the port network there is a clique corresponding to each ship, consisting of all ports it visited. The high efficiency of both networks is interesting, however, as it indicates that messages or goods can be passed along relatively short paths between pairs of ships or ports. The greater efficiency in the ship network suggests it is even easier in some ways to pass goods between ships than it is between ports. This is likely to be true for many networks involving both mobile entities and fixed positions, a factor that should influence the way we think about "control" in two-mode networks such as this one. Having influence in the mobile aspect of a network may be much more valuable than on fixed positions.

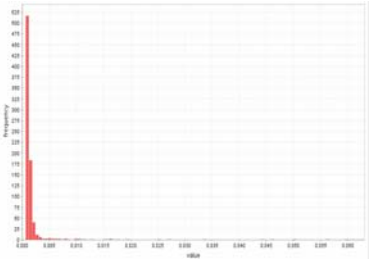
A growing trend in network analysis research is to characterize graphs according to several archetypical structures including hierarchical, cellular, core-periphery, and scale-free networks. Both derived graphs feature cliques which could be considered to form a cellular structure, but this should be ignored as an artifact of the process we used to generate them. This excluded, the only archetypes which stand out visually in our graphs are a possible core-periphery structure in the Shared Port network and a clearer, 2-core system in the Shared Ship network. Ports on the periphery may be of interest as supporting more varying commercial and social standards than the tightly connected inner portion of the merchant marine network. The histograms in figure 6 show that the distributions of node degree in these networks do not match well-known distributions, with the exception that it might be possible to fit a power-law degree distribution to the Shared Port or unimoded Stopped At networks. Power-law degree distributions are associated with scale-free networks and the small-world property. However, arguments of this sort are questionable as recent work has shown that exponential is an "attractor" distribution which many relationship sampling schemes tend to produce regardless of network structure (Bonacich, 2006; Airoldi & Carley, 2006).



A. Shared Ship



B. Shared Port



C. Stopped At

Figure 6. Histograms of Node Degree

4.3 Node Level Properties

In this section, we apply 3 primary measures of node centrality, each associated with a different type of significance within the network structure.

The **degree centrality** of a node is proportional to the number of edges leading into or out of it. Nodes with high degree centrality are typically "leaders" in their domain: they must be, to attract so many connections, and the immediate network around them is large and therefore rich in resources. They also experience a heavy workload since relationships normally require effort to maintain (in social networks, this is often referred to as *cognitive demand*).

The **eigenvalue centrality** of a node is similar to degree centrality, but is additionally affected by the degree of a node's neighbors, the degree of their neighbors, and so on. A node with high eigenvalue centrality is not only well connected but is surrounded by other well connected nodes. The measure differentiates between anomalously strong members of weak communities and elite members of a well-connected core (CITE).

The **betweenness centrality** of a node is proportional to the number of times it appears on the shortest path between two other nodes. High-betweenness nodes fill important, boundary spanning positions in the network. These nodes can have significant power as gate-keepers, since routing around them is expensive or impossible. In a social context, they also experience unique stresses by having to conform to the standards of multiple communities evolving in relative isolation. (CITE)

These measures are frequently highly correlated within a given network, so nodes for which some measures are anticorrelated are of special interest as "specialists" with the graph. For example, a node with high betweenness but low degree might be an especially efficient gatekeeper between two disconnected network regions.

Table 3(a) outlines top scorers in all 3 measures as applied to the Ship → Ship network. All measures are normalized against the maximum possible score. Ships appearing for more than one measure have had their names colored to aide identification. A striking characteristic of the results in this network is the lack of correlation between high scorers in the three measures.

Although the two top scorers in degree and eigenvalue centrality are constant, they do not even appear in betweenness centrality. This suggests that the ships be prolific travelers of well worn routes: they ports with many ships within a well connected community, but are not extraordinary in their itinerary so as to provide a potential bridge. The high scorers in betweenness centrality, by contrast, must cover unusual routes so that they are the only ship linking disparate regions.

Rank	Total Degree		Eigenvalue Centrality		Betweenness Centrality	
	Ship ID	Score	Ship ID	Score	Ship ID	Score
1	DDERFG	0.3155	DDERFG	0.0071	TPLIFQ	0.0580
2	7JUE7M	0.3155	7JUE7M	0.0071	7A9QL8	0.0480
3	6TTI00	0.3075	6TAPD8	0.0063	70FE3O	0.0426
4	7EDCPN	0.2941	AHOH1G	0.0061	7EDCPN	0.0411
5	6DT4H8	0.2687	70JB3O	0.0061	6T83A8	0.0390

(A) Top Centrality Nodes for the Shared Port (Ship → Ship) Network

Rank	Total Degree		Eigenvalue Centrality		Betweenness Centrality	
	Place ID	Score	Place ID	Score	Place ID	Score
1	33	0.4600	30	0.0485	33	0.1186
2	30	0.4200	33	0.0472	30	0.0917
3	32	0.3600	32	0.0394	0	0.0905
4	0	0.3400	0	0.0338	49	0.0824
5	26	0.3000	25	0.0336	18	0.0660

(B) Top Centrality Nodes for the Shared Ship (Place → Place) Network

Table 3. Node-Level Centrality Scores

Table 3 (b) outlines the same measures for the Place → Place network. In stark contrast, there is a very high level of correlation, with only a few locations occupying top slots across the board. In this network, the same locations are well connected, have well connected neighbors, and are essential stops on all nearby routes. The exception to this generalization is a higher level of variation in the betweenness scores. Examining locations "49" and "18", we find that they are ports near the Eastern bounds of the data, occupying positions between some outlying clusters and the main data. These are points of interest for efficiently observing and controlling outlying portions of the network.

5. Intervention Analysis

We now consider a potential intervention in the merchant marine network, in which ports will be requested to implement new security policies requiring increased inspection of all ships coming through them. We model the data being captured as A) being intrusive to gather and B) having a long "shelf life", so that it is unnecessary to gather the data repetitively for the same ship in a short time span. A good example – and one which relates to future CASOS merchant marine study -- is collecting detailed crew information from stopped vessels. If we are tracking long term patterns in crew movements between ships, it may be unnecessary to investigate every member of a ship at each place he stops.

The scenario described above is intended to create a need for efficiency by setting up a tension between thoroughness of surveillance and a reasonable level of effort on both port security and docked ships under an expensive and intrusive policy. One way to manage this balance is to select a subset of ports which will implement the new protocol. Ideally, the ports would be chosen such that A) a minimal number of ports are used (to save overhead on training personnel to enact the policy), B) the maximum number of distinct ships pass through ports enacting the policies (maximizing data acquisition), and C) the minimum number of total searches must be conducted (minimizing redundant searches). We model this formally by saying that for a set *P* of ports enacting the new protocol, the utility of the policy is

$$U(P) = \left| \bigcup_{p \in P} Ships(p) \right| - |P| c_p - \sum_{p \in P} |Ships(p)| c_s$$

Where *Ships(p)* is the set of ships visiting the port in a given timeframe, c_p is the cost of an additional port implementing the policy (where 1 unit is the value of a piece of information), and c_s is the cost of surveying each ship, in the same units. Alternatively, when cost estimates are unavailable, as they are in our case, we can examine the relative efficiencies of two ways of selecting ports by graphing the benefit (number of unique ships observed) against the imposition (total number of interventions required).

Under this framework, we can compare two different policies for selecting ports. A naïve approach might be to conduct surveillance at only the busiest ports, where the most ships dock. We can accomplish this by taking the highest eigenvalue centrality ports in this Ship → Place network. An alternative approach might be to pick high degree ports in the Place → Place network, since these presumably would receive a diverse array of ships from many neighboring locations. If efficiency were the primary concern, one might choose ports with high betweenness and low degree, as those boundary spanning locations might be more likely to witness *distinct* sets of ships.

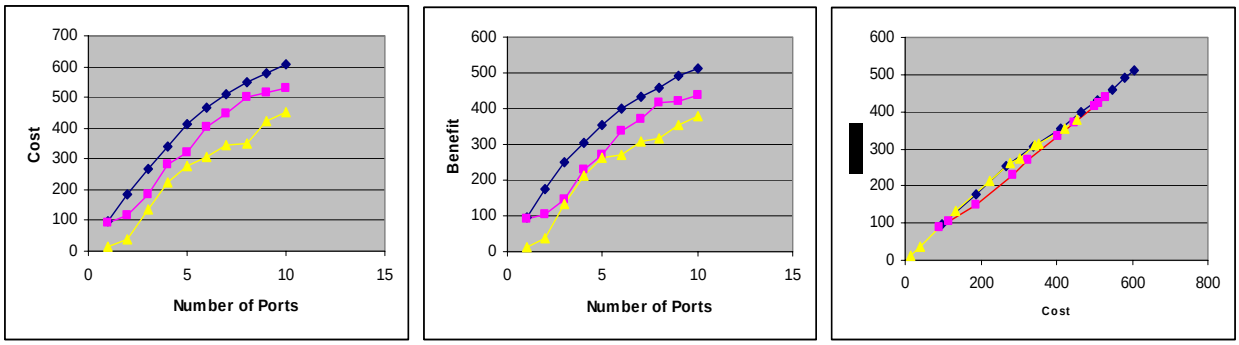


Figure 7. Cost/Benefit analysis of Surveillance Policies:

Busiest (blue), Highest Port → Degree (red), and High Betweenness (yellow)

The series of graphs in figure X show the relative strengths and weaknesses of these three proposed policies for 1 to 10 ports. The first policy – to conduct surveillance of the busiest ports – is the most aggressive. Each additional port increases dramatically both the cost and benefit created by this policy. The most conservative approach is the betweenness based. Surprisingly, these approaches are equally efficient. Both dominate the approach using Port → Port degree, which causes many more redundant observations while achieving more or less the same benefit

as the betweenness policy. The fact that all three policies performed so similarly is partially a consequence of the pattern we identified in the previous section, that in the Port → Port network there is little distinction between varieties of central role.

6. Discussion and Future Work

The main goal of this study was to provide proof-of-concept for an analysis framework that could, in a principled manner, 1) extract relational information from spatial data, 2) apply network analysis to find patterns in these relationships, and 3) model and advise policies regarding interventions. Multi-stage studies of this nature face many problems not found in experiments with a smaller scope: noise created by translating between models overwhelms signal, and false signals are injected as artifacts of the transition process. This experiment was successful in that at every stage of the experiment, patterns were identified that had meaningful interpretation in the original context. First, we were able to show that almost all of the locations of interest identified by our clustering algorithm corresponded to known ports, and that most of the remaining were substantially supported by the data. At the network level, we were able to identify significant differences in overall architecture between the ship → ship and port → port graphs, including a greater level of "specialization" (distinct types of central roles) in the ship → ship network. Finally, in our intervention analysis, we proposed a model of limited surveillance and showed that the network enforced a strict tradeoff between depth of surveillance and number of redundant observations.

A drawback of the breadth of this study is that the analysis conducted at each stage was necessarily cursory and could use further refinement. The clustering algorithm applied in our spatial analysis required human supervision and gave some bad results due to its inability to ignore outlying data. We are currently doing a much deeper study of this problem and plan to replace this algorithm in our pipeline with a much richer, probabilistic approach. A central goal is to be able to extract more behavioral information than simple locations of interest – we would like to extract information about types of activities and temporal relationships.

The network analysis presented here used the best studied, most accepted array of network and node measures. One direction for expansion is into newer techniques, such as modern grouping algorithms or measures intended for multimode matrices. As with the spatial data, another component we would like to incorporate is over-time analysis examining the evolution of the network throughout the timeframe.

Perhaps the most compelling area for future work that we touched upon was intervention analysis. One way to augment the intervention model presented here would be to compute an optimal allocation of ports and compare this to the heuristic policies which we discussed. However, a more serious issue is our implicit assumption that a policy like this can be based on historical data *with no expectation that implementing the policy will change agent behavior.* This is a frequent assumption in intervention modeling literature, but ignores the significant adaptability of human agents. Revisiting our model, "how can we best allocate surveillance based on today's traffic patterns?" might be a poorer question than "how can we allocate surveillance so that it is difficult for a deviant agent and well informed to route around?" Answering the second question requires not only descriptive analysis of patterns in data, but

inference of goals underlying agent behavior. CASOS is currently working on data-based game theoretic approaches to exactly this variety of question.

Appendix

ORA

The Organizational Risk Analyzer (ORA) is a comprehensive platform for the analysis of multi-mode networks. With over 5 years of development, it features many standard network analysis algorithms and a number of experimental measures being designed at the Computational Analysis of Social and Organizational Systems (CASOS) Lab at Carnegie Mellon University. It reads and records files in the extensible DyNetML format, and features standard network visualizers, GIS visualizers, over-time analysis tools, and more. For more information, including publications, see: http://www.casos.cs.cmu.edu/projects/ora/

The Merchant Marine Visualizer

Although not utilized in the body of this report, the Merchant Marine Time Tracker visualization is an important part of our current work on temporal analysis of spatial networks. It visualizes agents or other entities moving across locations over time. For instance Figure 1 shows five people moving from one city to another. Each location is shown as a column of nodes. Each agent is shown as a colored arrow. The arrows point to the location each agent was recorded at for each time period.

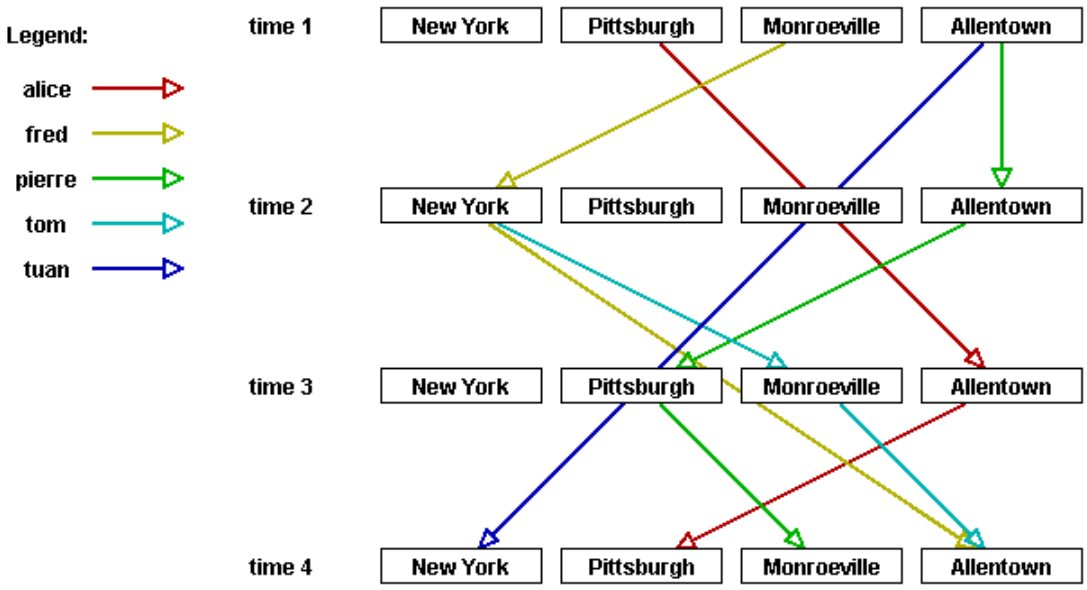


Figure 8. ORA MMV Trails Visualizer

The input consists of one meta matrix, with two entity sets and one graph per time step. One entity set represents the agents or entities to be tracked. The other entity set represents the locations the agents are moving between. Each graph maps the agent to location relationship for that time period.

The visualization can be accessed within ORA by running the Merchant Marine report. To get the visualization, load a DyNetML file with the attributes described above. Then click Analysis->Generate Reports. Then select Merchant Marine as the report type from the drop-down box, and selected the meta matrix you want to run the report on. Then click Next. On the second page select the entities that you want to track over time. (They will appear as arrows in the visualization). Then click Finish. The report should be generated and appear as a new page in your web browser.

ORA Geographic Information System

The Geographic Information System in ORA is a visualization tool for analysis of social networks with geospatial meta-data. Many real world datasets have geospatial distribution information for agents, knowledge or resource. Furthermore, it has been known that organizational performances, such as shared situation awareness, are dependent on the physical proximity of agents in an organization. The visualization of a network on a physical map and accompanying analysis methods/measures are important to comprehend the status of the organization and to predict the performance in the future. GIS in ORA supports the visualization and the simple analyses of a network loaded on the ORA interface.

GIS in ORA requires latitude and longitude information of each the node distribution. A user can specify this information in DyNetML, an xml file format for the presentation of social networks in an organization. An example DyNetML entry follows:

```
<node id="L2" title="CampLocation">
<properties>
<property name="latitude" type="double" value="70.0"/>
<property name="longitude" type="double" value="-135.0"/>
</properties>
</node>
```

Alternatively, a user can specify the longitude and the latitude of a location node and link the location node to the other nodes on the location.

For the MMV project, we created a set of hypothetical social networks of agents with the location information about where the vessels and marines are. The locations of the entities are chosen from the harbors located at countries in Pacific-Rim, US, UK, etc. Though the set is only the synthesized data, it is a data that resembles to the real dataset. Therefore, we use this dataset to validate our analysis and visualization methods. Furthermore, our dataset has evolving synthesized networks corresponding to one year period, which gives us a chance to show the evolution of networks and agent/vessel movements and interactions. The below images are the visualization of synthesized networks corresponding to four quarters in a year.

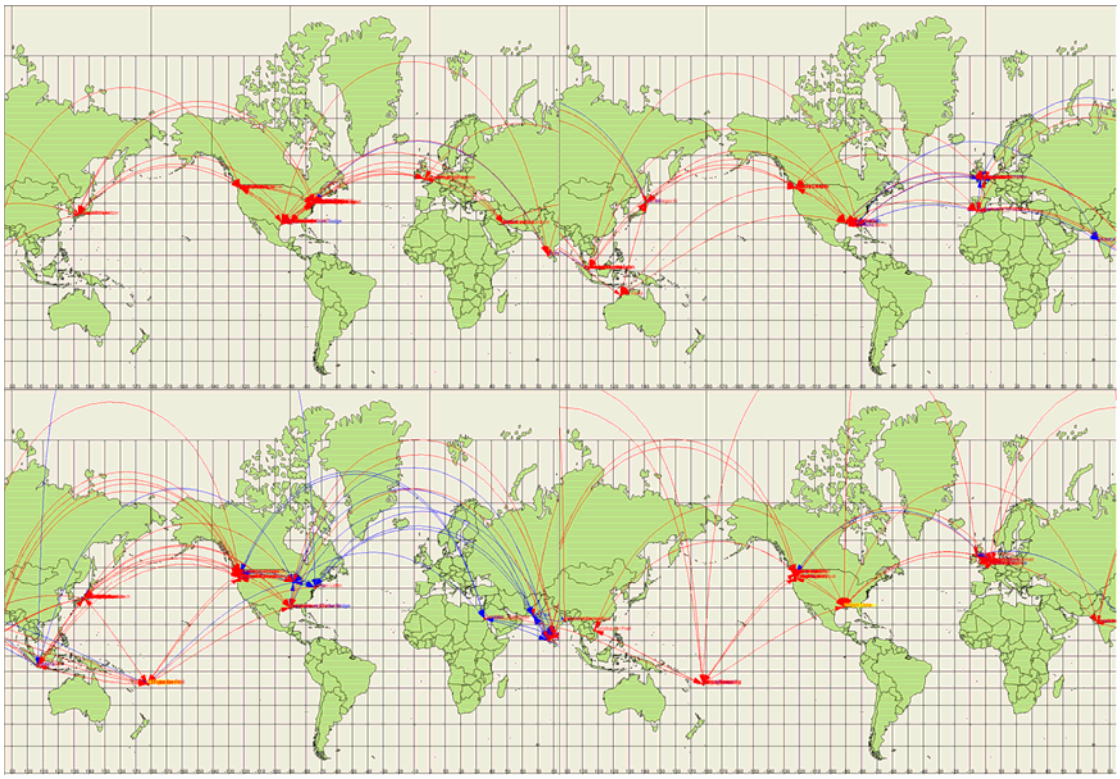


Figure 9. ORA GIS Visualizer

GIS in ORA is developed based on an open source GIS package, OpenMap built by BBN technologies.

A second technology we are leveraging is Google Earth (http://earth.google.com). Google Earth is a free tool that accesses a huge online database of satellite imagery and map data. ORA exports Google's KML markup language, allowing GIS visualizations such as the one below.

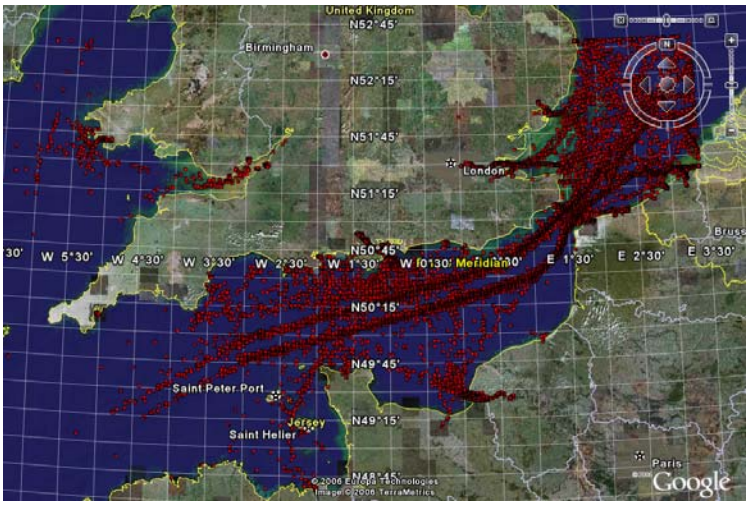


Figure 10. Google Earth Visualization